Research Paper | Information Technology | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

RESEARCH ON DARK DATA ANALYSIS TO REDUCE DATA COMPLEXITY IN BIG DATA

Bansari Trivedi 1 | Dr. Gokulnath K. [2]

[1] M. Tech Student, Information Technology Department, Parul Institute of Engineering and Technology, Waghodia, Vadodara, India.
[2] HOD, I.T. Dept. Parul University.

ABSTRACT

Big data is a large amount of data which is hard to handle by traditional systems. It requires new structures, algorithms and techniques. As data increases, dark data also increases. In such way there is one portion of data within a main data source which is not in regular use but it can help in decision making and to retrieve the data. This portion is known as “Dark Data”. Dark data is generally in ideal state. The first use and defining of the term "dark data" appears to be by the consulting company Gartner. Dark data is acquired through various operational sources but not used in any manner to derive insights or for decision making. It is subset of Big Data. Usually each big data sets consists average 80% dark data of whole data set. There are two ways to view the importance of dark data. One view is that unanalyzed data contains undiscovered, important insights and represents an opportunity lost. The other view is that unanalyzed data, if not handled well, can result in a lot of problems such as legal and security problems. In this phase solution for side effects of dark data on whole data set is introduced. Dark data is important part of Big Data. But it is in ideal state so it may cause load on system and processes. So it is important to find solution such that dark data should remain same and also can't affect rest of data.

KEYWORDS: Dark Data, Chunking, Classification, Big Data.

1. INTRODUCTION

Dark data is data which is usually stored in data source and remains unused but generally used in decision making. For example, a company may collect data on how users use its products, internal statistics about software development processes, and website visits. However, a large portion of the collected data is never even analyzed.

The following categories of unstructured data usually are considered dark data:

- Customer Information
- Log Files
- Previous Employee Information
- Raw Survey Data
- Financial Statements
- Email Correspondences
- Account Information
- Notes or Presentations
- Old Versions of Relevant Documents

Suppose there is a telecommunication company who is generating big data per day. Now it has customer details and communication records as a dark data. From those details company will analyze a behavior of customer and find out their interests and comforts. In other hand company can also find out that in which type of situation they can earn more?

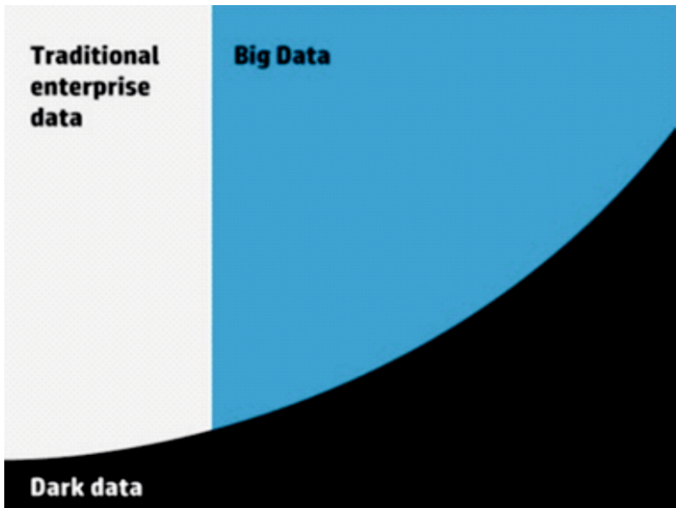


Fig. 1.1 Growth of dark data

As per one survey there is average 80% dark data of total size of data. Dark data is mostly exists in Big Data structure. Traditional databases have lower chances to generate dark data.

2. LITERATURE REVIEW

In this comprehensive survey study, we review the background of data deduplication and the differences between data deduplication and traditional data compression. We also comprehensively study the state-of-the-art work on data deduplication, classifying them into six general categories based on data deduplication workflow, and then create taxonomy for each category, which provides insights into the pros and cons of existing solutions. [1]

I-sieve is also useful thing for this. It is a high performance inline deduplication system for use in cloud storage. We design novel index tables to satisfy the I-sieve architecture, since it is a bridge between frontend and backend systems. [2]

These data provide opportunities that allow businesses across all industries to gain real-time business insights. The use of cloud services to store, process, and analyze data has been available for sometimes it has changed the context of information technology and has turned the promises of the on-demand service model into reality. In this study, we presented a review on the rise of big data in cloud computing. [3]

In the current deduplication scenario, the content-based chunking methods provide good throughput as well as decrease the space utilization. While working and managing the multimedia files like images, videos, or audio, the content-aware chunking methods are accepted mostly. [4]

Recent times have witnessed a growing use of Virtualization and Cluster environment in order to optimally handle and use resources. Major problem with existing management systems is that they are developed for native systems, which poses serious problems when need for changing underlying environment arises. [5]

3. PERFORMANCE EVALUATION

In today's system dark data is about 80% of total data. It may cause more load on system and also affects processing power as well as processing speed. Advantages dark data are decision making and retrieving data are easy with help of dark data.

In other hand as a disadvantage system load, duplication, data complexity, memory and power consumption etc. can be consider. Deduplication can be done..

- Post process: after store the data
- In-line: real time deduplication

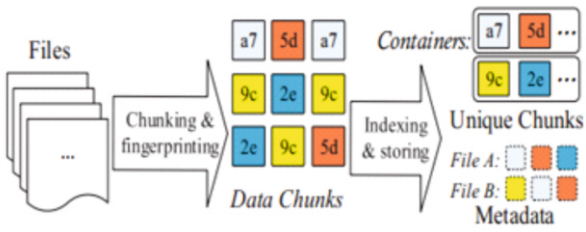
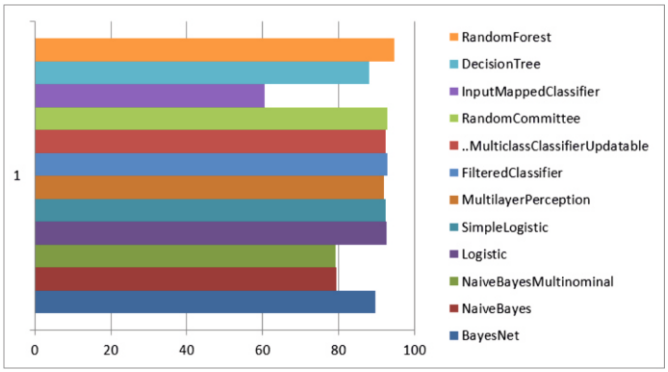


Fig. 3.1 Process of deduplication [1]

Deduplication means reduce number of replications by removing replicated data or by give two different identities to two same data. Deduplication helps in reducing complexity, memory consumption, decision latency etc. in data source. As shown in figure-3.1 in deduplication has 3 phases. In first phase data sets with duplicated data items are selected. In second phase those data sets are divided in chunks means manageable parts of main data. Then these chunks will be index and analyze to detect duplicated data. At the end in third phase duplicated data will be separated in form of metadata. And remaining data will be store in main source.

For complete classification of spam based dataset



X=Values in percentage

Y=Name of method

Here, proposed work is to manage and store duplicated or dark data and divide them into clusters. For this chunking method can use. Those clustered data can retrieve temporarily on cloud for further required process.

In this way user can also separate metadata as a dark data and data required for backup also can separate from main data source. This will reduce load and memory as well as power consumption of main data handling server. Complexity will reduce.

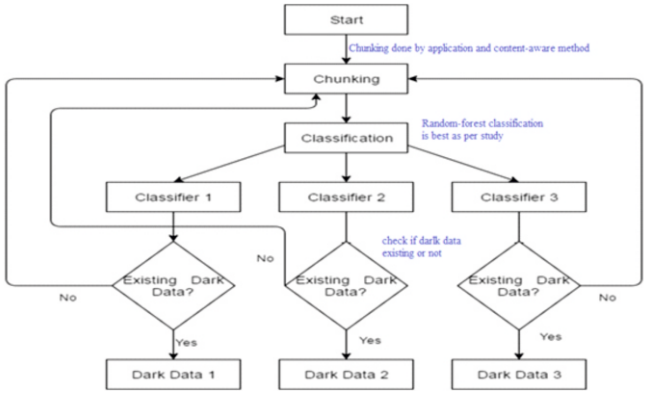


Fig. 3.2 From Data to Dark Data Flowchart

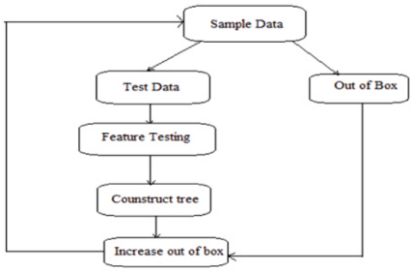


Fig. 3.3 Flowchart for Random-forest classification

4. CONCLUSION

Here, we can separate dark data from main data such that we can use dark data only during decision making and rest whole data remains unaffected. We can also deduplicate data to reduce complexities occurring due to replicated data. Separated data or data sets can be store on cloud storage. Their analyzed results can also store in cloud so we can use it on requirements rather than use whole data or data set. Dark data is most general and useful part of big data but it is necessary to process dark data to reduce its side effects on data source.

REFERENCES

Research papers:

[1] "A Comprehensive Study of the Past, Present, and Future of Data Deduplication", Wan Xia, Member, IEEE, Hong Jiang, Fellow, IEEE, Dan Feng, Member, IEEE, Fred Douglis, Senior Member, IEEE, Philip Shilane, Yu Hua, Senior Member, IEEE, Min Fu, Yucheng Zhang, and Yukun Zhou, MANUSCRIPT ID 0203-REG-2015-PIEEE (BASE PAPER)

[2] "I-sieve: An Inline High Performance Deduplication System Used in Cloud Storage", Jibin Wang, Zhigang Zhao, Zhaogang Xu, Hu Zhang, Liang Li, and Ying Guo, TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214ll03/11llpp17-27 Volume 20, Number 1, February 2015

[3] "The rise of "big data" on cloud computing: Review and open research issues", Ibrahim Abaker Targio Hashem, Ibrar Yaqoob , Nor Badrul Anuar , Salimah Mokhtar , Abdullah Gani, Samee Ullah Khan , Information Systems 47 (2015) 98–115

[4] "Understanding the Dark Side of Big Data Clusters: An Analysis beyond Failures", Andrea Rosa, Lydia Y. Chen, Walter Binder, 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks

[5] "Distributed Virtualization Manager for KVM Based Cluster", Mr. Uchit Gandhi, Mr. Mitul Modi, Ms. Mitali Raval, Mr. Paavan Maniar, Dr. Narendra Patel, Prof Kirti Sharma, Procedia Computer Science 79 (2016) 182 – 189, ScienceDirect

[6] "Data Model for Big Data in Cloud Environment", Imran Khan, S. K. Naqvi, Mansaf Alam, S. N. A Rizvi

[7] "Study of Chunking Algorithm in Data Deduplication", A.Venish and K. Siva Sankar, Springer India 2016

Websites:

[8] http://www.kdnuggets.com/webcasts/index.html

[9] https://en.wikipedia.org/wiki/Random_forest